

Übungen zu Ökonometrie - Blatt 5

(Abgabe: Donnerstag, 08.06.2011, vor den Übungen)

Aufgabe 1 (17 Punkte)

Auf der Homepage der Vorlesung befindet sich die Datei `swisslabor.dat`, die die Ergebnisse einer Befragung von 872 Haushalten über Gesundheit in der Schweiz enthält. Dabei wurden Daten über folgende Größen erhoben.

- Teilnahme: Ist die Person erwerbstätig?
- Alter: Alter in Jahrzehnten (Jahre geteilt durch 10)
- Bildung: Anzahl der Jahre der Berufsausbildung
- JKinder: Anzahl der Kleinkinder (unter 7 Jahre)
- AKinder: Anzahl der älteren Kinder (über 7 Jahre)
- Herkunft: Ist die Person Ausländer (also kein(e) Schweizer(in))?

Teilnahme	Alter	Bildung	JKinder	AKinder	Herkunft
no	3	8	1	1	no
yes	4.5	8	0	1	no
⋮	⋮	⋮	⋮	⋮	⋮

- (a) Betrachte das Logit-Modell und regressiere die binäre Variable "Teilnahme" (Zielvariable) auf alle übrigen (erklärenden) Variablen. Interpretiere den Output der R-Funktion `summary()` nach Anwendung auf das Ergebnis von `glm()`. (4)
- (b) Erweitere das Logit-Modell aus Aufgabe (a), indem als weitere erklärende Variable das Quadrat des Alters mit aufgenommen wird und interpretiere wie in (a) den R-Output. (4)
- Hinweis:
Als erster Parameter in `glm()` muss hier `Teilnahme~1+Alter+Bildung+JKinder+AKinder+Herkunft+I(Alter^2)` angegeben werden.
- (c) Welches der beiden Modelle aus (a) und (b) ist im Sinne des AIC-Kriteriums besser? (1)
- (d) Was bedeutet im R-Output `Number of Fisher Scoring iterations`? (3)
- (e) Teste mit Hilfe des Likelihood-Quotiententests aus der Vorlesung, ob die Regressionskoeffizienten β_3 (Bildung), β_4 (JKinder), β_5 (AKinder) und β_6 (Herkunft) im Logitmodell aus Aufgabe (a) gleich 0 sind (Nullhypothese: $H_0 : \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$). Interpretiere das Testergebnis. (2)

Vorgehensweise in R:

Durchführung einer Logit-Regression unter H_0 :

```
> glm.swiss1 <- glm(Teilnahme~1+Alter, ...)
```

Durchführung einer Logit-Regression mit allen erklärenden Variablen:

```
> glm.swiss2 <- glm(Teilnahme~1+Alter+Bildung+JKinder+AKinder+Herkunft, ...)
```

Anwendung des Likelihood-Quotiententests:

```
> anova(glm.swiss1, glm.swiss2, test="Chisq")
```

Der p-Wert steht in der Spalte `P(>|Chi|)`.

- (f) Teste analog zu Teilaufgabe (e) die Hypothese $H_0 : \beta_6 = 0$ vs. $H_1 : \beta_6 \neq 0$ im Logitmodell und interpretiere das Ergebnis. (3)

Aufgabe 2 (10 Punkte)

Auf der Homepage der Vorlesung befindet sich die Datei `motorins.dat` mit Daten über Autoversicherungsschäden in Schweden für das Jahr 1977. Folgende Variablen sind im Datensatz enthalten:

- Kilometer: Gefahrene Kilometer pro Jahr
1: weniger als 1000
2: zwischen 1000 und 15000
3: zwischen 15000 und 20000
4: zwischen 20000 und 25000
5: mehr als 25000
- Zone: Geographische Zone
1: Stockholm Göteborg, Malmö mit Umgebungen
2: andere große Städte mit Umgebungen
3: kleinere Städte mit Umgebungen in Südschweden
4: Ländliche Gegenden in Südschweden
5: Kleinere Städte mit Umgebungen in Nordschweden
6: Ländliche Gegenden in Nordschweden
7: Gotland
- Bonus: Bonus-Klasse des Autofahrers
- Marke: 1–8 repräsentieren verschiedene typische Automarken. Alle anderen Marken sind in Klasse 9 zusammengefasst.
- Versicherte: Anzahl der Versicherten (pro Versicherungsjahr)
- Schaeden: Anzahl der Schäden

Kilometer	Zone	Bonus	Marke	Versicherte	Schaeden
1	1	1	1	455.13	108
1	1	1	2	69.17	19
⋮	⋮	⋮	⋮	⋮	⋮

- (a) Betrachte das verallgemeinerte lineare Regressionsmodell mit der Poisson-Verteilung für die Zielvariable "Schaeden" und natürlicher Linkfunktion (Hinweis: `family=poisson(link="log")`). Interpretiere den Output der R-Funktion `summary()` nach Anwendung auf das Ergebnis von `glm()`. (4)
- (b) Teste, ob mindestens eine der erklärenden Variablen Einfluss auf die Anzahl der Schäden hat. (3)
- (c) Teste, ob die Anzahl der Schäden von der geographischen Zone abhängt. (3)